

Réunion 9 février 2026

- Survey on RAG Meeting LLM: Towards Retrieval-Augmented Large Language Models [Fan et al., 2023]
- Towards Trustworthy Retrieval Augmented Generation for Large Language Models: A Survey [Ni et al., 2025]
- Administratif et Organisationnel

Extraction - Types

- **Types d'Extracteur**
 - **Sparse retrieval** : au niveau des mots (TF-IDF, BM25)
 - pas entraîné
 - grosse dépendance à la qualité de la BDD et requête
 - **Dense retrieval** : plonger dans un espace
 - 1ere couche de générateur pour projeter
 - encodeur spécifique ([Contriever](#) - [\[Izacard et al. 2022\]](#))
 - bi-encodeur (Dense Passage Retrieval - [\[Karpukhin et al. 2020\]](#))

Extraction - Granularité

- **Granularité de d'Extraction** : taille de découpe des extraits
 - **chunk retrieval** - le plus répandu
 - **token retrieval** - plus fin mais plus lourd en calcul
 - **entity retrieval** - perspective du langage (Entities as Experts - [\[Févry et al. 2020\]](#))

Extraction - Amélioration pré-retrieval

- **Pre-Retrieval Enhancement** : réduire la charge du retriever en *complétant la requête*
 - **query expansion** - générer des docs fictifs, compléter la query avec les infos pertinentes ~> désambiguïser (Query2Doc - [\[Wang et al. 2023\]](#))
 - **query rewrite** - reformuler l'entrée à l'aide données supplémentaires (Rewrite-Retrieve-Read - [\[Ma et al. 2023\]](#))
 - **query augmentation** - compléter l'entrée avec une réponse sans données extraites (ReFeed - [\[Yu et al. 2023\]](#))

Extraction - Amélioration post-retrieval

- **Post-Retrieval Enhancement** : réduire la charge sur l'extracteur en *affinant les informations extraites*
 - **ajustement du classement** des extraits (PRCA - [Yang et al. 2023])
 - **réduction du bruit** dans les extraits
 - assembler **plusieurs extracteurs** (R²G : Retrieve Rerank Generate - [Glass et al., 2022])
 - **ajouter la query** aux données
 - utiliser des **reasoning paths** (Search-in-the-Chain - [Xu et al. 2024])
 - **compression/synthèse** des (gros) extraits

Survey on RAG Meeting LLM: Towards Retrieval-Augmented Large Language Models [Fan et al., 2023]

Extraction - Base de données

Nature et source des données différentes

- **closed source** - locale
- **domain specific**
- **open-source** - utilisation de moteurs de recherche public

Survey on RAG Meeting LLM: Towards Retrieval-Augmented Large Language Models [Fan et al., 2023]

Générateurs - Familles

Différents modèles de langues ~> Dépend de la tâche tierce

- **Parameter accessible** - peut être affiné
- **Parameter inaccessible** - via API sans accès aux paramètres

Intégration - Stratégies

- **Input Layer Integration** - Combiner Requête et Extraits en entrée du Générateur
 - tous les documents **en même temps** (In-Context RALM - [Ram et al., 2023])
 - un document **à la fois** (ATLAS - [Izacard et al. 2023], REPLUG reproduction - [Shi et al., 2024])
- **Output Layer Integration** - Fournir les données et le texte généré séparément (ou petit traitement pour unir) (ReFeed - [Yu et al. 2023])
- **Intermediate Layer Integration** - intégrer les documents extraits au milieu de la génération (du modèle)

Intégration - Fréquence et Nécessité

Extraction non pertinente $\implies$ hallucination / bruit

=> Définir des critères de recours ou non au Retrieval

- **Coté RA-LLM**

- token spécifique
- prompt itéré
- modèles annexes entraînés pour déclencher
- définition de logit déclencheurs

- **Coté RAG "traditionnels" :**

fréquence fixée

- one-time
- every-n-token
- every-token

Entraînement - Stratégies

Affiner les composants sur la tâche tierce

- **Training free**
 - **prompt engineering based method** : injection des données dans le prompt
 - **retrieval guided token generation method** : calibrer la génération à l'aide des données
- **Independent Training** : hors pipeline
- **Sequential Training** : le 2nd au sein du pipeline (2 ordres possibles)
- **Join Training** : les deux en même temps dans le pipeline

Perspectives de recherche

- RA-LLM multilingue
- RA-LLM multimodal
- Gestion de la qualité des données
 - Construction de base de données avec fiabilité plus régulière que Wikipedia
 - Extraction qui filtre les informations peu fiables
- Trustworthy RA-LLM

Trustworthy RAG

- **Reliability** : cptmt adapté systématique, quantification de l'incertitude, généralisabilité
- **Privacy** : protection des données utilisateur
- **Explanability** : processus de décision transparent, génération des sorties compréhensible par utilisateur
- **Fairness** : abs de biais
- **Accountability** : capacité à justifier le respect des réglementations
- **Safety** : Contenu qui ne heurte pas (attaques adverses, acteur malicieux, DB poisoning,...)

Trustworthy RAG - Explainability

Explication à l'**Extraction** (*peu de recherche ~> IR*)

- **post-hoc** : feature attribution, génératif -> *model agnostic?*
- **axiomatic strategies** [\[Völske et al. 2021\]](#), [\[Anand et al. 2023\]](#)
- **probing strategie** : analyse des paramètres et encodage, sensibilité aux variations de paramètres
- **self-interpretable design** : by-design

Trustworthy RAG - Explainability

Explicabilité à la génération:

- **post-hoc**
 - basé sur la perturbation (RAG-Ex [\[Sudhi et al., 2024\]](#))
 - par construction d'exemple contrefactuel
- **ante-hoc** - explications intégrées au modèle
 - ~> permet le perfectionnement du raisonnement
 - reasoning on graph (RoG - [\[Luo et al., 2024\]](#))

Trustworthy RAG - Explainability Metrics

Métriques classiques en XAI

- **Fidélité**
- **Stabilité** : régularité des résultats
-> divergence de Kullbach-Lieber,
correlation,...

Données d'évaluation : pas de dataset canonique

- adaptation/ complétion d'un dataset
- dataset spécifique à la tâche traitée

Métriques pour le RAG

- **Significance** : capacité à exploiter les infos importantes (F1-score, Mean Reciprocal Rank)
- **Plausibilité** (évaluation humaine)

Trustworthy RAG - Explainability Perspectives

- Intégration des Knowledge Graph (hybride)
- Méthodes d'explication qui prennent en compte **tous les composants** du modèle
- Trade-off Explicabilité et performance aussi
- Abs de métrique d'évaluation pr RAG
- DB Poisoning à gérer

Premières lectures - Benchmarks

- BIRCO [[Wang et al. 2024](#)]
 - LLM-based Retrieval
 - tâches et thèmes variables
- Bright [Su et al., 2025]
 - reasoning intensive tasks
 - thèmes divers (scientifique)
- *BEIR*
 - plus accessible ...

Questions suscitées:

- Qu'attend-on du modèle (en terme de compétences)?
 - Réflexion nécessaire pour trouver les docs?
 - Extraction basées sur LLM?
- Est-on capable de faire poser des questions supplémentaires préliminaires à un Chatbot?

Administratif et organisationnel

- [Conf : 2e Symposium Sante mentale et IA](#) des présentations intéressantes pour ARTESU
 - Agents conversationnels **Jeu 12/02 14h-15h**
 - Farzaneh Lashgari (University of Beira Interior) *SENTINEL-LLM: Suicide Ensemble-Based Text Intelligence and Natural Language Evaluation Through Large Language Models* (30 min) **Ven 13/02 9h35**
- Cours Machine Learning Données Séquentielles